Research Report

Small-Sample Equating by the Circle-Arc Method

Samuel A. Livingston

Sooyeon Kim

July 2008

ETS RR-08-39

Small-Sample Equating by the Circle-Arc Method

Samuel A. Livingston and Sooyeon Kim
ETS, Princeton, NJ

July 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS' constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This paper suggests two new, related methods for estimating a test-score equating relationship from small samples of test takers. These methods do not require the estimated equating transformation to be linear. Instead, they constrain the estimated equating curve to pass through 2 prespecified end-points and a middle point determined from the data. Some preliminary results indicate that these methods outperform mean equating and other methods used for equating in small samples.

Key word: Equating, small samples, curvilinearity, strong models, mean equating

The Problem

Often it is necessary to equate scores on a new form of a test taken by a very small number of test takers: 30, 20, or even fewer. Equating test scores on the basis of such a small sample of test-takers is likely to produce results that will not generalize well to other groups of test takers. As the statisticians responsible for the equating of the scores, we cannot make the problem go away by claiming that the small group of test takers whose new-form scores we can observe are the entire target population (i.e., the population for which we want the equating to be correct). A test taker's reported score should not depend heavily on the particular group of test takers who happened to take the test at the same time that he or she did. We need to determine an equating relationship that will generalize—at least approximately—to other groups of test takers.

One common way that statisticians deal with the problem of small samples is to use "strong models" that reduce the number of parameters to be estimated from the data—in effect, substituting assumptions for data. In test score equating, the most common versions of this approach are linear equating and, using an even stronger model, "mean equating" (Kolen & Brennan, 2004, p. 125), which is essentially linear equating with a prespecified slope. Yet, when test forms differ substantially in difficulty, the equating relationship between them is typically not linear—not even approximately linear. A difficult form and an easy form, administered to the same group of test takers, will produce differently skewed distributions. The difficult form will spread out the scores of the higher ability test takers and bunch together the scores of the lower ability test takers. The easy form will do the opposite. Consequently, the slope of the equating transformation will not be the same for the weaker test takers as it is for the stronger test takers. A linear transformation, with its constant slope, cannot capture this aspect of the equating relationship. The inaccuracies tend to be greatest at the ends of the score distribution, where the linear equating transformation may extend beyond the range of scores possible on the reference form. We need a better method.

Circle-Arc Method 1

The method proposed here is based on an idea from Divgi (1987): to constrain the equating curve to pass through two prespecified end-points and an empirically determined middle point. In Divgi's method, the end-points were determined by the maximum and minimum possible scores on the test forms to be equated. The middle point was determined by the mean scores. The estimated equating transformation was a cubic polynomial passing through

those three points, with the slope of the curve at the middle point equal to the ratio of the standard deviations.

The method proposed here is somewhat different, but it also constrains the equating curve to pass through two prespecified end-points and an empirically determined middle point. The upper end-point of the curve is determined by the maximum possible score on each form. The lower end-point of the curve is determined by the lowest meaningful score on each form. On a multiple-choice test, this score would typically be the *chance score*—the expected score for a test-taker who answers without reading the questions. (It would be possible to choose some other point as the lower end-point of the curve, if there were a reason to do so.) A third point on the curve is determined from the data by equating at one point in the middle of the score distribution. If those three points happen to lie on a straight line, that line is the estimated equating curve. If the three points do not lie on a straight line, they determine an arc of a circle. In Circle-Arc Method 1, the circle arc that passes through these three points becomes the estimated equating curve. To extend the equating function below the lower end-point of the curve (i.e., below the lowest meaningful score), that point is connected linearly to the point representing the minimum possible scores on the two forms.

Figure 1 illustrates Circle-Arc Method 1. In this hypothetical example, a new form of a test consisting of 20 four-choice items is being equated to a reference form also consisting of 20 four-choice items. The three points that determine the circle arc are indicated by the square boxes. The upper end-point of the curve, determined by the maximum possible scores, is specified to be (20, 20). The lower end-point of the curve, determined by the chance scores, is specified to be (5, 5). The middle point has been determined from the data to be (12, 14), which implies that the new form is more difficult than the reference form. The circle that includes these three points is centered at (40, -15) and has a radius of $\sqrt{1625} \doteq 40.31$. To extend the equating transformation below the lower end-point of the curve, which represents the lowest *meaningful* scores, that point has been connected linearly to the point representing the lowest *possible* scores, in this case (0, 0).

Formulas for Method 1

Label the coordinates of the three points to be connected by a circle-arc as (x_1, y_1), (x_2, y_2), and (x_3, y_3). The lower end-point of the curve, specified in advance, is (x_1, y_1); the

upper end-point, also specified in advance, is (x_3, y_3); and the middle point, determined by the data, is (x_2, y_2). Let *r* represent the radius of the circle, and label the coordinates of its center (x_c, y_c). The equation of the circle is $(X - x_c)^2 + (Y - y_c)^2 = r^2$ or, equivalently, $|Y - y_c| = \sqrt{r^2 - (X - x_c)^2}$.

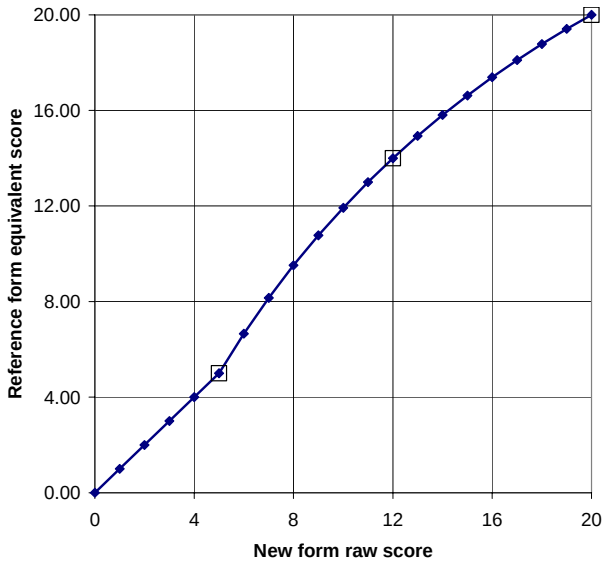

Figure 1. **Illustration of Circle-Arc Method 1.**

If the new form is harder than the reference form, the middle point will lie above the line connecting the lower and upper points, so that the center of the circle will be below the arc. For all points (X,Y) on the arc, $Y > y_c$, so that $|Y - y_c| = Y - y_c$, and the formula for the arc is

$$Y = y_c + \sqrt{r^2 - (X - x_c)^2}. \quad (1)$$

If the new form is easier than the reference form, the middle point will lie below the line connecting the lower and upper points, so that the center of the circle will be above the arc. For all points (X,Y) on the arc, $Y < y_c$, so that $|Y - y_c| = y_c - Y$, and the formula for the arc is

$$Y = y_c - \sqrt{r^2 - (X - x_c)^2}. \quad (2)$$

A simple decision rule is to use Equation 1 if $y_2 > y_c$ and Equation 2 if $y_2 < y_c$.

The formulas for x_c and y_c are a bit cumbersome:

$$x_c = \frac{(x_1^2 + y_1^2)(y_3 - y_2) + (x_2^2 + y_2^2)(y_1 - y_3) + (x_3^2 + y_3^2)(y_2 - y_1)}{2\left[x_1(y_3 - y_2) + x_2(y_1 - y_3) + x_3(y_2 - y_1)\right]} \; ; \tag{3}$$

$$y_c = \frac{(x_1^2 + y_1^2)(x_3 - x_2) + (x_2^2 + y_2^2)(x_1 - x_3) + (x_3^2 + y_3^2)(x_2 - x_1)}{2\left[y_1(x_3 - x_2) + y_2(x_1 - x_3) + y_3(x_2 - x_1)\right]} \; ; \tag{4}$$

but the formula for r^2 is simply

$$r^2 = (x_1 - x_c)^2 + (y_1 - y_c)^2 \tag{5}$$

These values can then be substituted into Equation 1 or Equation 2.

Circle-Arc Method 2

The curves produced by Circle-Arc Method 1 differ in shape—slightly, but systematically—from the curves that equipercentile equating commonly produces. Fortunately, there is another way to use the three points to determine an estimated equating curve. This alternative procedure is a fairly simple modification of Circle-Arc Method 1, and it produces curves that are more similar to those produced by equipercentile equating in large groups. We call this modification Circle-Arc Method 2.

In Circle-Arc Method 2, the function that estimates the equating curve is divided into two components, a linear component and a curvilinear component. The linear component is simply the line connecting the two points specified as the end-points of the curve. We will denote this line by $L(x)$. Algebraically,

$$L(x) = y_1 + \frac{y_3 - y_1}{x_3 - x_1}(x - x_1) \, . \tag{6}$$

(If the new form and the reference form are alike in test length and item format, $L(x)$ will be the identity line.)

The curvilinear component is the difference between the estimated equating function and the line $L(x)$. We will denote it by y^*, so that

$$y^* = y - L(x)\,. \tag{7}$$

Figures 3–6 illustrate the procedure for Method 2, using the same numerical example as for Method 1. Figure 3 shows the three points that determine the equating curve, exactly as in Method 1. Figure 2 also shows the linear component $L(x)$, which is the straight line connecting the two prespecified points.

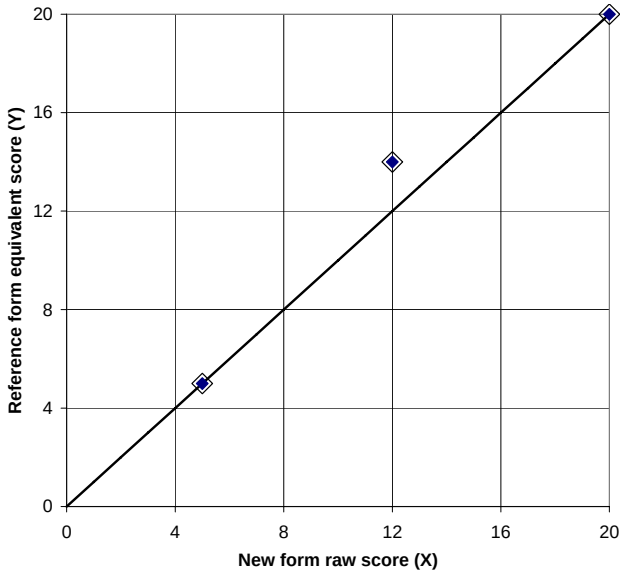

Figure 2. **Circle-Arc Method 2: determining three points on the equating curve.**

In Figure 3, these three points are transformed by subtracting $L(x)$, leaving the curvilinear component y^*. Because the two end-points of the curve are on the line $L(x)$, this step transforms their y values to zero:

$$y_1^* = y_3^* = 0\,. \tag{8}$$

The height of the transformed middle point (x_2, y_2^*), near the bottom of Figure 3, is equal to the vertical distance of the original middle point (x_2, y_2) above the line $L(x)$ connecting the two prespecified end-points. This distance depends on what the small-sample equating indicates about the difficulty of the new form and the reference form. If the equating indicates that the new form is harder than the reference form, the middle point will be above the line connecting the end-points, and y_2^* will be positive. If the equating indicates that the new form

is easier than the reference form, the middle point will be below the line connecting the end-points, and y_2^* will be negative.

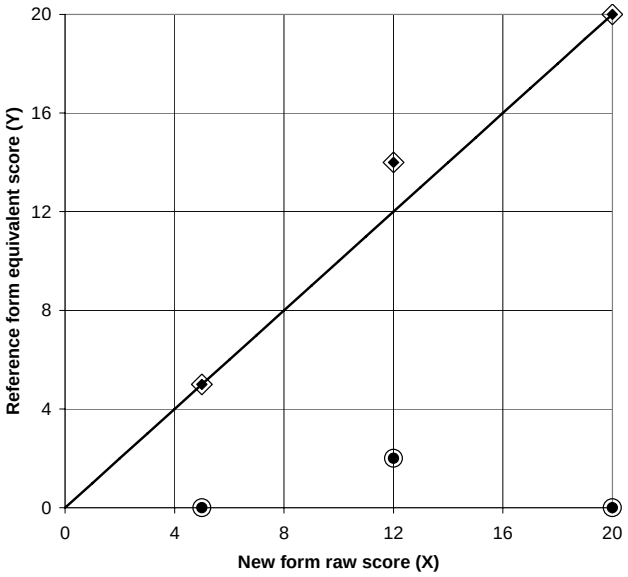

Figure 3. **Circle-Arc Method 2: transforming the three points.**

In Figure 4, the three transformed points are used to determine a circle arc, and this arc determines the value of *y** at each new-form raw score. The formulas for determining the circle arc are exactly the same as those for Method 1, except that *y** is substituted for *y*. In this example, those formulas determine the center of the circle to be (12.5, -13) and its radius to be $\sqrt{225.25} \doteq 15.01$.

Finally, in Figure 5, the curve for *y** is retransformed back to the score scale for the reference form, by adding in the height of the line *L*(*x*). The resulting curve is an estimate of the equating function for scores between the specified end-points. The estimated equating function is not a circle arc; it is the sum of a circle arc and a line with a positive slope of approximately 1.

Determining the Middle Point

If the equating design[1] is a single-group, counterbalanced, or equivalent-groups design, the middle point (x_2, y_2) can be determined by equating the mean score on the new form directly to the mean score on the reference form.

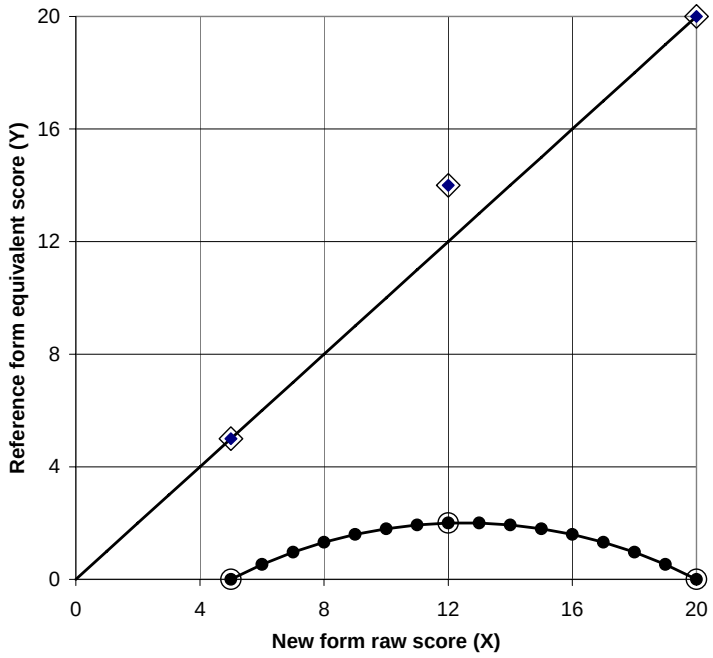

Figure 4. **Circle-Arc Method 2: fitting the circle arc to the three transformed points.**

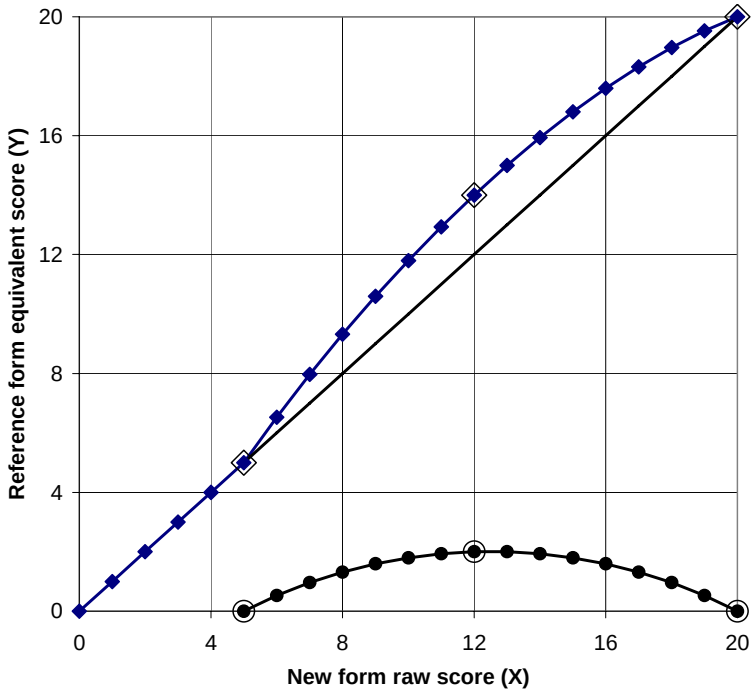

Figure 5. **Circle-Arc Method 2: retransforming the points determined by the circle arc.**

In an anchor equating design, the middle point can be determined by chained linear equating at the mean score of the smaller group of test takers—typically, the group taking the new form. The chained linear equating formula for the reference-form score *y* corresponding to new-form raw score *x*, is

$$y = m_{YB} + \frac{s_{YB}}{s_{VB}}\left(m_{VA} - m_{VB}\right) + \frac{s_{YB}}{s_{VB}}\frac{s_{VA}}{s_{XA}}\left(x - m_{XA}\right) , \tag{8}$$

where *m* and *s* indicate the means and standard deviations, *A* and *B* indicate the test-taker groups taking the new form *X* and reference form *Y*, and *V* indicates the anchor score. This formula yields an equated score *y* for any value of *x*, even if *x* is not actually a possible score on the new form. Therefore, we can choose $x_2 = m_{XA}$, so that the equating formula simplifies to

$$y_2 = m_{YB} + \frac{s_{YB}}{s_{VB}}\left(m_{VA} - m_{VB}\right) . \tag{9}$$

This method of determining the middle point requires only two pieces of information from the new-form group: their mean score on the test and their mean score on the anchor. It requires four pieces of information from the reference-form group—the means and standard deviations of their scores on the test and on the anchor. In most cases, if the new form and the reference form have been taken by substantially different numbers of test takers, the reference form will have been taken by the larger group. However, if the group taking the reference form is the smaller group—by enough to matter—it would be better to determine the middle point (x_2, y_2) by choosing y_2 to be the mean score of the reference-form group and using chained linear equating to determine x_2, the corresponding score on the new form.

A Tryout of the Two Methods

To see how the two methods would work in practice, we conducted a small-scale resampling study. The criterion equating for this study was a common-item equating of two forms of a 107-item, four-option multiple-choice test. Each form had been taken by more than 6,400 test takers. Table 1 shows statistics describing their scores on the anchor and on the full test. The groups taking the two forms were about equally strong; their mean scores on the common items differed by only about 0.03 standard deviations. However, on the full test, the

mean score of the group taking the new form was about 0.36 standard deviations lower than that of the group taking the reference form, indicating that the new form was substantially more difficult.

Table 1

Statistical Comparison of the Groups Taking the new Form and Reference Form

Statistic	New-form test takers	Reference-form test takers
Number of test takers	6,426	6,489
Anchor score mean	30.60	30.46
Anchor score *SD*	4.96	5.09
Standardized mean difference	0.03	
Test score mean	73.62	77.47
Test score *SD*	10.51	10.83
Standardized mean difference	-0.36	
Correlation of test and anchor	0.91	0.92

The criterion equating was a chained equipercentile equating of presmoothed score distributions, computed in the full group of 6,400+ test takers taking each form. The presmoothing was a log-linear bivariate smoothing, applied to the joint distribution of the total score and the anchor score in each group of test takers, preserving five univariate moments of each marginal distribution (i.e., of the total score and of the anchor score) and one cross-product moment. The procedure for the study consisted of repeatedly drawing pairs of small samples from the group of test takers who took the new form and from the group who took the reference form, equating the test forms by using the data from the small samples, and comparing the results with those of the (large-group) criterion equating.

We wanted to compare the two circle-arc methods not only with each other, but also with other equating methods likely to be used in small-sample situations. We included three linear equating methods: (a) the Tucker method, (b) the Levine observed-score method, and (c) the chained linear method. We also included two versions of mean equating: the version presented by Kolen and Brennan (2004, p. 125, equation 4.78) and a version based on the chained approach. The results of the two versions were nearly identical, and only those of Kolen and Brennan's version will be reported. Since the criterion equating was a chained equipercentile equating of presmoothed distributions, we also included that method. However, in applying it to

the small-sample data, we used a stronger smoothing model, preserving only three univariate moments of each marginal distribution and one cross-product moment. (Preserving the third moment of each distribution is necessary to capture the curvilinearity in the equating relationship.)

The resampling study consisted of 200 replications of the following procedure:

1. Draw a random sample of n_x test takers who took the new form and a random sample of n_y test takers who took the reference form.

2. In those samples of test takers, equate the new form to the reference form by all the selected equating methods.

In choosing the sample sizes to use in the study, we wanted to use numbers that are fairly typical of those encountered when a new form of a small-volume test is introduced. Typically, the reference form has been administered more than once, so we specified that the reference-form sample would include 3 times as many test takers as the new-form sample. We decided to draw samples of 25 test takers taking the new form and 75 taking the reference form.

Figure 6 shows how much the small-sample equated scores differed from those produced by the population equating, as indicated by the root-mean-squared difference (RMSD) over the 200 replications. The units of both the horizontal and vertical scales are raw-score points (i.e., correct answers), but the vertical scale is magnified, to show the differences among the small-sample methods. Although the range of possible raw scores extends from 0 to 107, only 5 of the 6,426 test takers taking the new form had scores lower than 38, and only 4 had scores higher than 100. The horizontal scale of the graph extends from raw score 30 (just above the chance score of 26.75) to 107 (the maximum possible score). The 10th, 25th, 50th, 75th, and 90th percentiles of the new-form raw-score distribution are indicated by "X" symbols on the horizontal scale.

Figure 6 shows that the differences between methods in accuracy were small for raw scores near the median of the distribution but large for scores far from the median. In these small samples, the methods based on strong assumptions (i.e., mean equating and the two circle-arc methods) clearly outperformed those based on weaker assumptions (i.e., linear equating and equipercentile equating). The most accurate method, overall, was Circle-Arc Method 2. In every part of the score range, its RMSD was either the smallest or very nearly the smallest. Its

advantage over mean equating was substantial below the 10th percentile and above the 90th percentile.

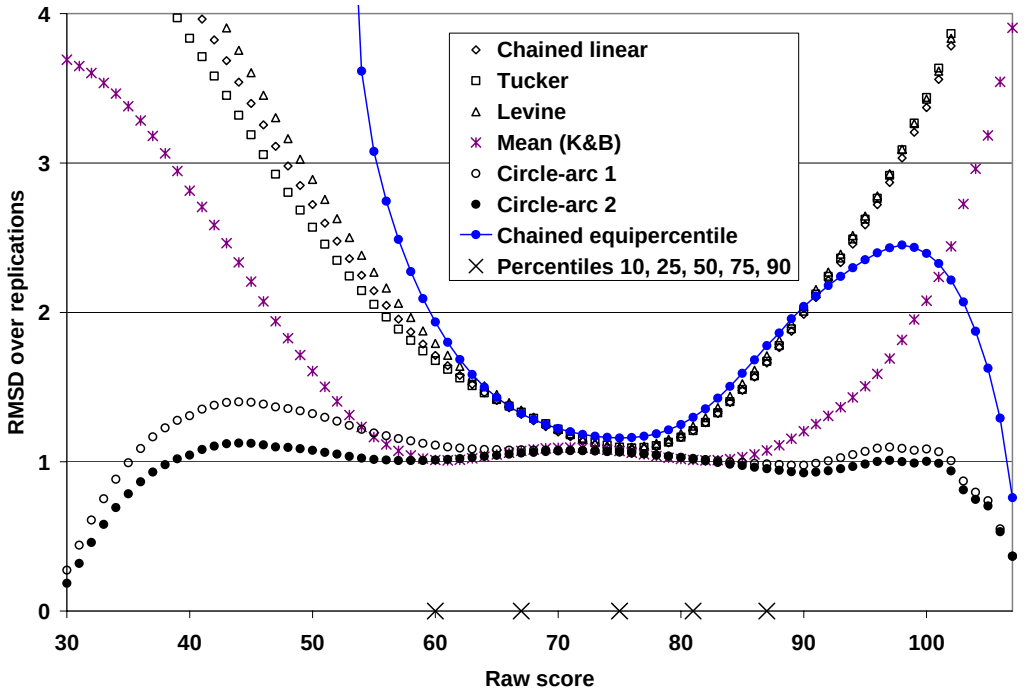

Figure 6. **Root-mean-squared difference (RMSD), over 200 replications, of small-sample equatings from population equating.**

Figures 7 and 8 decompose the RMSD into its two components. One component, labeled "bias over replications," is the mean of the differences between the small-sample equated scores and the criterion equating. The other, labeled "SD over replications," is the standard deviation of those differences.

Figure 7 shows that in the middle of the distribution, all the small-sample methods had a negative bias in equating this particular pair of test forms. That is, they produced equated scores that were too low, making too small an adjustment for the greater difficulty of the new form. This bias was smallest for the chained equipercentile method. Above the 75th percentile, all the linear methods (including mean equating) showed a positive bias, increasing for scores farther from the median of the distribution. This result was inevitable, because of the curvilinearity in the criterion equating. The two circle-arc methods showed a negative bias in this portion of the score range, smaller for Method 2 than for Method 1. The chained equipercentile method showed

less bias than the other methods, except in the lowest portions of the score range, well below the 10th percentile.

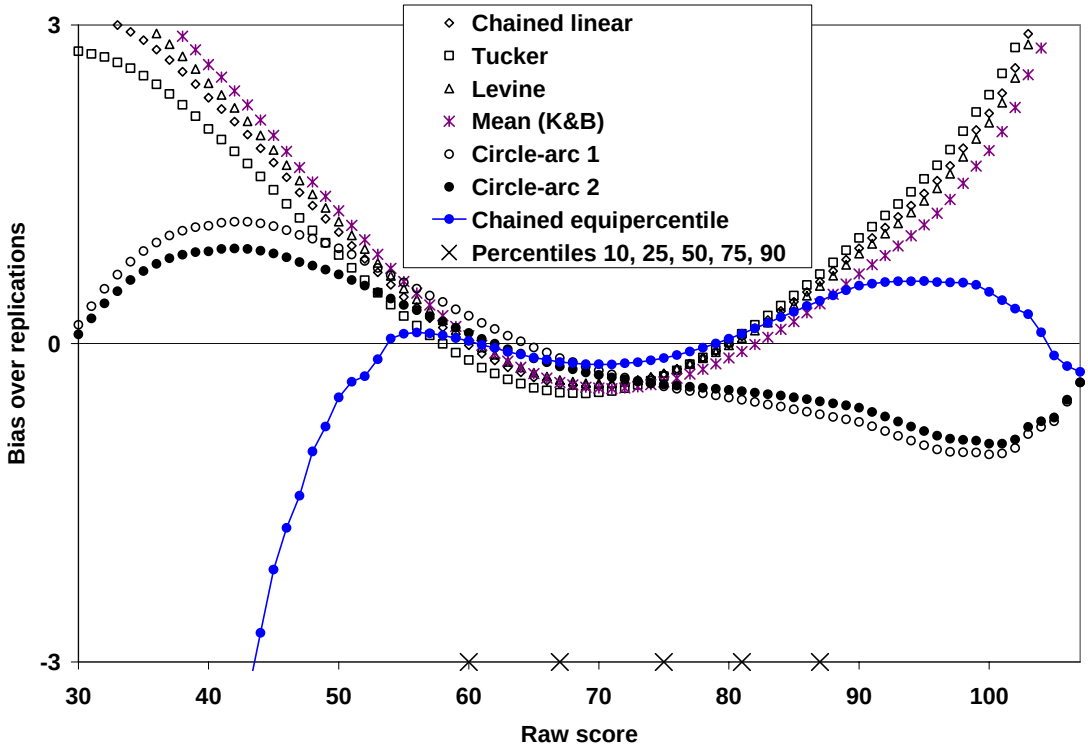

Figure 7. **Bias (over 200 replications) in small-sample equating by each method.**

Figure 8 clearly shows the main reason for the greater accuracy of the methods based on strong assumptions: smaller sampling variability. For mean equating, the standard deviation over replications is constant over the score range. For circle-arc equating, the standard deviation over replications decreases to zero as the raw score approaches one or the other of the two prespecified data points.

Limitations of the Method

Like other strong models, circle-arc equating cannot reproduce certain types of equating transformations. If one of the two test forms to be equated has more items of medium difficulty, while the other has more hard items and more easy items, the equipercentile equating function will tend to be somewhat S-shaped. Neither circle-arc method can model this more complex curvature. The circle-arc estimate of the equating function in this situation will tend to be close to the identity. Mean equating also would produce a line close to the identity, but linear equating

would produce a line with a different slope. If the slope of the line were estimated accurately (not necessarily a safe assumption if the samples are small), the linear equating would tend to be more accurate than circle-arc equating through much of the score range but less accurate near the end-points.

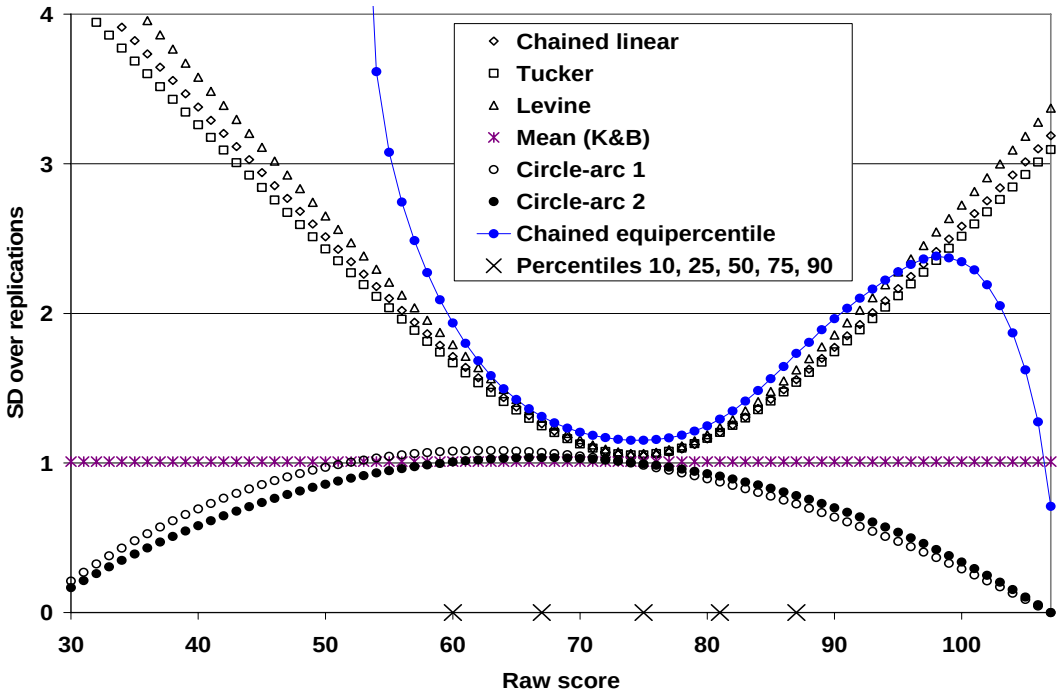

Figure 8. **Standard deviation (over 200 replications) of equated scores by each method.**

It is possible to imagine a situation in which mean equating would produce a more accurate estimate than the circle-arc approach. Suppose the difficulty difference between the two test forms came from the replacement of one or more very easy items by items so difficult that almost no test takers answered them correctly. In this case, the equating transformation would be close to that produced by mean equating. Although such a scenario is possible, it does not seem likely to occur often in practice.

What Next?

Circle-arc equating appears to be a strong candidate to replace mean equating as the method of choice for small-sample data, if the results of this preliminary tryout generalize to other pairs of forms and to other tests. Circle-arc equating also may be preferable to linear

equating in situations where test forms differ in difficulty and the samples are too small for equipercentile equating. Probably the best way to determine the extent of its usefulness (i.e., the conditions under which it produces better results than other methods) would be to conduct a series of resampling studies like the one described in this paper. The factors to be investigated would include the equating design and the sample size. The sample sizes to investigate would be different for the different equating designs. An equivalent-groups design requires much larger samples than an anchor-test design, which requires larger samples than a counterbalanced design. In evaluating the results at each score level, it might be necessary to look more closely at the distribution, over replications, of the differences between small-sample equatings and the criterion equating. The present paper focuses on only the mean and standard deviation of that distribution, but the frequency of large differences also may be important.

References

Divgi, D. R. (1987). *A stable curvilinear alternative to linear equating* (Rep. No. CRC 571). Alexandria, VA: Center for Naval Analyses.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Notes

[1] See Livingston (2004, pp. 27-35) for an explanation of equating designs.